“Document Image Compression”

Ph.D. Workplan

Submitted to

Swami Ramanand Teerth Marathwada University, Nanded

Submitted by

P. L. Paikrao
Research Scholar

Under the Guidance of

Dr. D. D. Doye
Professor
Department of Electronics and Telecommunication Engineering

SHRI GURU GOBIND SINGHJI INSTITUTE OF ENGINEERING AND TECHNOLOGY, VISHNUPURI, NANDED, MAHARASHTRA, INDIA-431606

Synopsis: Document Image Compression

1. A Brief Resume of the Proposed Problem:

Document image compression reduces the storage requirements for digitized documents in order to optimize the costly storage and networking resources. This work shall investigate several fundamental areas of the document image compression process. Methods for segmentation of document images are investigated along with techniques for determining various regions in it. If images can be segmented into regions like text and non-text; it is possible to select an effective compression model to reflect the region's contents. Study of existing methods and developing the ad-hoc method to segment the document image regions, including the machine learning and deep learning schemes shall be carried out and the comparative findings based on accuracy, logical segmentation achieved, and subsequent compression improvements shall be discussed. The compression system presented will consist of separate compression technique for various regions in document images, depending upon their suitability, as a function of reconstruction image quality.

2. Introduction:

Today, many problems that arise with printed material are not related to the production aspects of papermaking, but with the physical storage requirements that such vast volumes of paper records required. The storage requirements for digital images are several orders of magnitude lesser than for paper-based documents. Document image compression reduces the storage requirements for digitized documents by using characters as the fundamental unit of compression.

2.1 Digitization:

Digitization is the technique of conversion of paper-based document records such as books, thesis, certificates, legal, legacy documents, maps etc., into the digital form. The document images are the electronically scanned versions of document records available at the Government and other offices. The current processes used for digitization of old document records, do not employ any standard procedure for working with the different regions in such multimodal document images. Usage of optical character recognition for printed or handwritten text is limited, as the language support is inadequate. The other regions in the image like photographs, figures, graphs, equations,

watermarks, holograms, logos, etc. are not treated separately, that's why the storage and transmission of digitized documents images is a costly affair and hence compression of such document images is needed.

2.2 Need of compression

The quantity of the old records that need to be digitized in various departments of government and other organizations, will ranges over millions of pages. To store a document page as a black and white image of moderate size viz. 512 X 512 pixels, one needs approx. 35 KB of disk space and a color image with 1024 X 1024 pixel will require 3MB of storage.

Table 1.1 Space and Transmission cost of various images

Sr. No.	A document page	Page Size	Spectral Resolution Bits/Pixel	Uncompressed image size	Transmission BW in bits per second	Transmission Time
1	BW Image	512 x 512	1 bpp	32.768 kB	32-64 kbps	1min 2 sec
2	Grayscale Image	512 x 512	8 bpp	262 kB	2.1 Mbps	1 min 13 sec
3	Color Image	512 x 512	24 bpp	786 kB	6.29 Mbps	3 min 39 sec
4	Medical Image	2048 x 1680	12 bpp	5.16 MB	41.3 Mbps	23 min 54 sec
5	SHD Image	2048 x 1680	24 bpp	12.58 MB	100 Mbps	58 min 15 sec

The Table 1.1 show Space and Transmission cost of various images from Black and White (BW) images to Super High Definition (SHD) images. The parameters like disk space, transmission bandwidth, transmission time needed to store and transmit such digitized data is reported.

If digitization of millions of documents is to be carried out, then it requires scanning of each page generally with a standard resolution of 600 dpi (dot per inch). Storing millions of such scanned pages will require significantly large amount memory and this will lead to improper and inefficient use of storage resources and its access mechanism. Compressing these millions of pages will save huge disk storage requirement. So, for the proper utilization of storage, computing and network resources there is a need of an efficient compression technique that will compress the large amount of data images and in turn will require very lesser amount of disk space and also decoding technique to get original or approximate form by decoding the code previously used for compression.

So, it becomes a motivation behind this work to put forth a technique that will effectively encode the scanned images into their compressed form before storage and subsequently its decoding technique to view the image in its original form.

3. Review of Literature and Development in the Relevant Subject:

3.1 Image Compression:

The most common method for scanning and storing the document is JPEG and JPEG2000. But while converting to the image, it reduces the visual redundancy. The most important assumption understanding the image redundancy is that greater frequency components are less important than the lower frequency components. But, in case of official receipt sheets or the official handwritten documents, the higher frequency components degradation will completely blur the lines or its strokes. In general, image compression denotes the data reduction process by retaining the image information. Compressing the data consists of the key components which are considered are, channel bandwidths, storage requirements and quality maintenance. If these parameters are achieved, a proficient data compression technique may be developed, which will save storing space and also speed up the transmission time.

As discussed earlier, digital images occupy a large storage space and take longer time to transmit or download [2][5], compression becomes necessary to reduce this. Thus, Image compression techniques are used to identify data redundancy and irrelevance internally and then develop an alternate representation of an image that uses less storage space than the original size of the image. The reverse process for regaining the older image form called decompression is employed.

Image compression is considerably different than the compressing of raw binary data. Images consist of certain statistical properties that encoders can exploit, so the obtained result is less than optimum when general purpose compression programs are used [7].

The need to compress an image becomes obvious when the number of bits/image is computed, which results from usual sampling rates and quantization procedures. For example, the

image of 512 x 512 pixels/colour,8 bits/pixel, and with three colors will require 6 x10^6 bits approximately storage space. In consequence, the storage of very few images can too cause a problem.

3.2 Image Compression Principles

The common characteristic of many images is that the correlation is present in neighboring pixels, and hence, information is redundant. The primary task is to discover less correlated depiction of the image. The two elementary constituents of compression are irrelevancy and redundancy reduction. A decrease in redundancy focuses on the removing of duplication from the image signal source, while the component irrelevancy reduction neglects the parts of the source signal that cannot be noticed by signal receiver, viz., the Human Visual System-HVS[1].

According to one of the studies, redundancy of three types are identified [3],

a. Spatial Redundancy: Correlation between values of neighboring pixel.

b. Spectral Redundancy: Correlation between different spectral bands or color planes.

c. Temporal Redundancy: Correlation between neighboring frames in the sequence of images (mostly applied to video applications).

As the study focused on still images, three more types of image redundancies are explored. They are,

a. Coding Redundancy

b. Psycho-visual Redundancy

c. Inter-pixel Redundancy

Coding redundancy occurs when fewer code words are used. While Inter-pixel redundancy occurs from correlations within the pixels of images. And the psycho-visual redundancy is the result of the fact that there lies insensitive in the human eye in respect to some spatial frequencies [1].

3.3 Classification of Image Compression Techniques

The various ways of classifying compression techniques are mentioned below,

1. Lossy Compression: - Using this technique, some of the more delicate details of the image are sacrificed to save storage space or Bandwidth [1]. The reconstructed data is acceptable approximation of original data. These technique offers higher compression ratio than the techniques mentioned subsequently. Its applications are in compression of natural images, audio, video, etc.

2. Lossless compression: - This system aims to reduce the bit rate of compressed output of the image without distorting the image. After decompression, the bit stream is the same as the original bit stream [1]. There is no loss of information, reconstructed data is identical to original data. Its applications are in text, medical imaging, law forensics, military imagery, satellite imaging, etc.

3. Near-Lossless compression: There is always a tradeoff between compression ratio and quality of reconstructed image. To have sufficient quality of reconstruction with required compression ratio, for a particular application [8]. An intermediate approach called near-lossless compression is proposed. Here the difference between the original and reconstructed data is within user-specified amount called as maximum absolute distortion or peak of absolute error. Its applications are in medical images, hyper spectral images, videos, etc.

3.4 Document Image Compression

The image files can be saved in many formats, like PNG, BMP, JPEG, GIF, and TIFF. Particularly, a scanned image when stored on digital devices occupies a lot of space, and hence, compression of scanned images becomes necessary. Technological developments in the field of storing, processing, and visualizing data have made it possible to handle many documents in the form of digital images and make them available on digital platforms. This entire process has two primary concerns. The first one is the size of the document. A document stored in ASCII version can easily have size until 2–3 kB, whereas a scanned document may occupy memory between 500 kB and 2 MB. To store and transmit documents in image form, an efficient and effective

compression technique are needed. The second concern is about its readability. Most lossy compression techniques reduce the resolution, which might affect the readability of the document image. For best results, it is important that the document should be readable even after retrieval from high-level lossy compression.

Most of the times, document images are nothing but scanned versions, which are mostly rich in textual content. Informally, document image can be defined as images containing components, which resemble with the language symbols. In the 1970s, when firstly the use of FAX machines started, the need for document image compression was identified as an important issue. Various document compression techniques offer us a facility to store a large amount of data most efficiently. Most of them use signal compression techniques to eliminate redundancies. It transforms the source data into a stream, which signifies the data by a compact encoding. The basic characteristics of the transformation process give us scope for compressibility. Transform coding and vector quantization are the two commonly used methods. The transformation can be focused on preserving the specific properties of the document. While maintaining the document appearance, resolution reduction may leave the text portions unreadable. All the attempts to texturally encode a document image will fail, as the useful data in a document image is obviously in the form of high-gradient, high-contrast, locally non-uniform pixel patterns. While going for lossy compression, care must be taken to maintain the readability of the document. Ascher and Nagy firstly suggested an approach for document compression, and later Witten et al. formalized the same, they attempted to encode data redundancy at the symbol level. The document image components can be indexed in dynamic library, and their repeated appearance is referred to the first instance. This may be further enhanced the form of constituent patterns and residuals. This method is quite the same as the existing pattern matching and substitution techniques, and it provides the best way to identify and eliminate similarities in the image components. After that, the redundancies in symbol shape are located.

4. Objectives of Proposed Research and Proposed Hypothesis:

Document image compression is relatively a new in the field of digitization (all type of documents) and significant work has yet to be done in this regard. The research related to general image compression and document image compression is relatively different. For the various documents having variations in the Format, Layouts, Script directions, Scripts Symbols, Font Styles etc. there is unavailability of Optical Character Recognition (OCR) systems. OCR systems work better for printed text, but for multitone document images, these methods may not work properly. Also, prediction-based method compresses the image with preserving data blindly, but these techniques may not achieve higher compression ratio. So, they are not suitable for this purpose.

Typical documents in today's office are computer-generated, but even so, inevitably by different computers and software such that even their electronic formats are incompatible. Some include both formatted text and tables as well as handwritten entries. There are different sizes, from a business card to a large engineering drawing. Document analysis systems recognize types of documents, enable the extraction of their functional parts, and translate from one computer generated format to another. Automated mail-sorting machines to perform sorting and address recognition have been used for several decades, but there is the need to process more mail, more quickly, and more accurately. In a traditional library, loss of material, misfiling, limited numbers of each copy, and even degradation of materials are common problems, and may be improved by document analysis techniques. All these examples serve as applications ripe for the potential solutions of document image analysis.

Most of the image compression techniques are developed for uniform tone images. e.g. natural images, text images, medical images, satellite images etc. Document background is not given enough attention and generally neglected. Text and Text-like attributes in graphical region are not given enough attention. Considering the trade-off between compression ratio and quality of reconstructed image, the adopted techniques are not optimized for these metrics as may be done using the near lossy compression techniques.

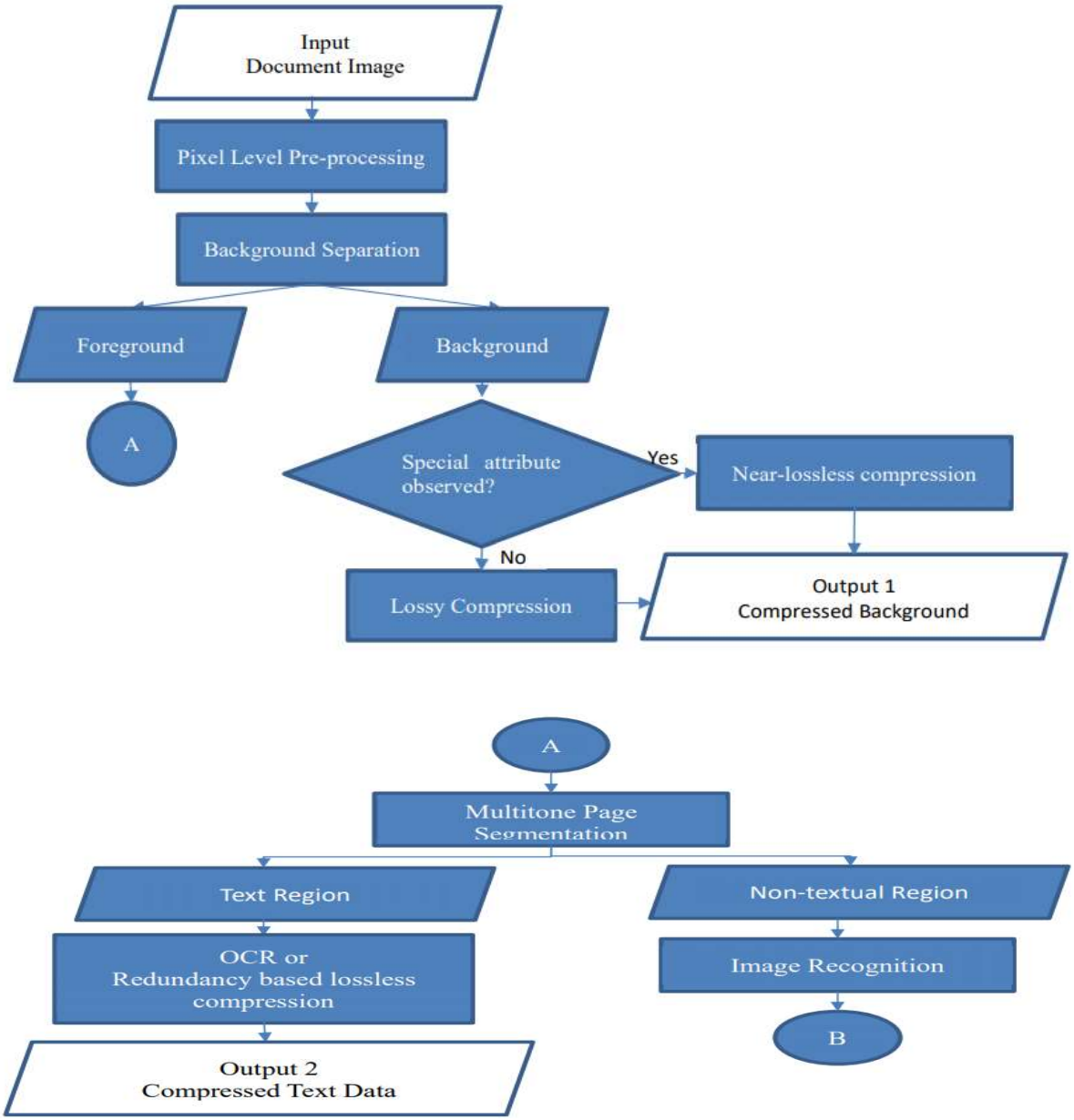
The objective of the research is to come up with the new and efficient image compression technique that will effectively compress the multimodal document images consisting of many

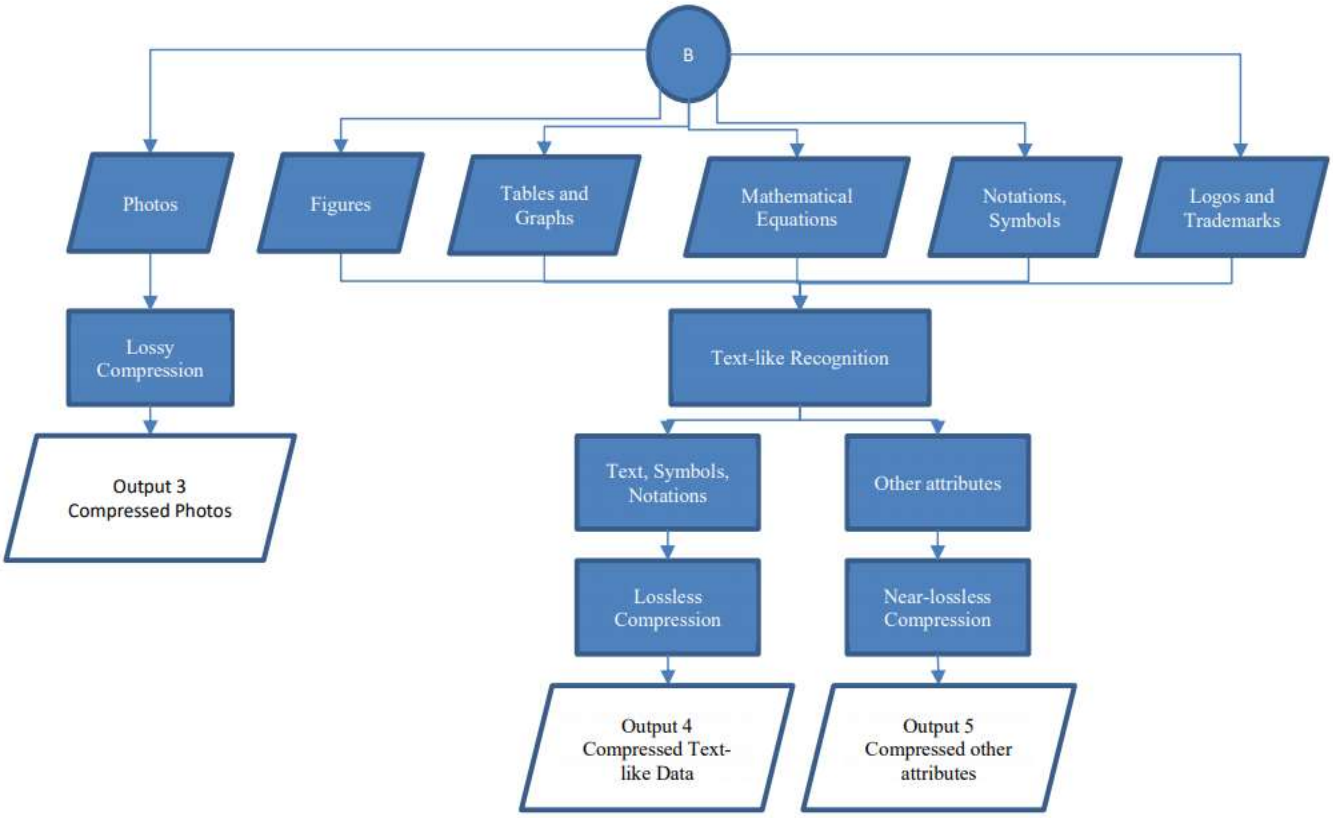
regions. This compression of the document images shall need lesser storage resources and network resources.

1. To study and compare different image compression techniques, algorithms.
2. To develop an efficient image compression algorithm suitable for document images, using different approaches that will optimize the compression ratio.
   - 2.1 Image smoothing techniques for preprocessing.
   - 2.2 Document Image Segmentation and background separation.
   - 2.3 Separate compression techniques for different regions of the document image e.g. background, text, text-like, and non-textual data.
   - 2.4 Exploring the suitability of near lossless compression techniques for graphical regions.
3. To analyse, validate, justify and recommend the proposed algorithm with respect to various parameters such as compression ratio, quality of reconstructed image and compression time etc.

5. Methodology to be used and adopted:

Compression gains can be achieved by identifying regions that contain textual and non-textual regions and applying a separate compression model to reflect the region's contents. In separated text, isolating unique characters, and storing them in a codebook using lossless compression techniques. On the other hand, applying near-lossless and lossy compression techniques on separated text-like and non-textual regions respectively shall result in optimum compression ratio. The following flow diagram shows the workflow of proposed methodology.





The proposed methodology Involve:

1. Preprocessing of document images using Edge Preserving Smoothing of Document images.
2. Lossless Compression for Text Regions using Burrows Wheeler Transform, Run Length Encoding, and Entropy coding schemes.
3. Near-lossless Compression for Figures, Tables, Graphs, etc. that contains text-like attributes.
4. Lossy Compression for Photographs.
5. Lossy or Lossless compression for Background.

6. Importance of study:

6.1 International Status:

With the advent of new technologies, traditional libraries are in need to go for Digital Library. In recent years, digital conversion of data has advanced rapidly, and this promotes some of the ordinary public to believe that entirety of essential data can be always be found on WWW - World Wide Web in the digital form. Moreover, digitization can be done for nearly every single format and medium of the library document like maps to manuscripts, and images to musical recordings. Digitization proves to be an excellent way to provide access to library resources, but current technology is not that adequate to archive the data appropriately and preserve it in the right form [4]. Storing these huge contents and sharing them on internet has become the need of the time, so the document image compression has worldwide importance in these regards.

6.2 National Status:

1. Nowadays, digital libraries are created for varied and different sectors, for example, science, culture, education, health, development, governance, etc. Automation of the library provides easy access to such collections using hi-tech library directory such as "Online Public Access Catalog (OPAC)". Moreover, the digital library is quite different from the traditional library in terms of allowing users for online access to the electronic data of text documents as well as their associated images.

2. In India, Karnataka State Archives stands to be a rich storehouse among all other states for ancient documents and manuscripts as old as 200 years. Further, this State Archive is also the repository to lakhs of printed historical documents of immense value and significance for historians, scholars, academicians, public and students. This state archive also possesses Old Government Files of Pre-independent era and even the Secretariat files which reflects the type of administration of that Karnataka state has undertaken over the period.
3. On the similar ground, Directorate of Archives of Maharashtra Government in 2016, also took steps to digitize old Books, Manuscripts, and other important government records. The tender issued for this stated that Archives Department of Maharashtra state plans to

commence the work of computerization or the digitization of around 15crores pages comprising historical printed records, manuscripts, and rare books.

4. Digi lockers: Digilocker is a flagship initiative of Ministry of Electronics & IT (MeitY) under Digital India Programme. It aims to 'Digital Empowerment' of citizens by facilitating them with Digital Document Wallet, that may be used to store, access anytime, anywhere and provide their authentic documents on internet. The newly machine generated documents as well as old records are Digitized in this Programme. So, minimizing the human efforts and intervention during digitizing the old documents this work may contribute.

7. Proposed Work Plan with Tentative Chapter Scheme

Chapter 1: Introduction

Chapter 2: Literature Review

Chapter 3: Document Image Pre-processing Techniques

Chapter 4: Document Image Segmentation

Chapter 5: Lossless Compression for Textual Region

Chapter 6: Lossy Compression for Photographs

Chapter 7: Near-lossless Compression for Text-like Regions

Chapter 8: Results, Discussions and Conclusions

8. Year wise plan of work and targets to be achieved

Year 1	Year 2	Year 3
Performing the Literature Survey to identify the Gaps in Literature	Study and Implementation of Lossless Compression for Textual Region in Document Images	Development of Complete System that combines and all the compressed outputs together
Study and Implementation of Document Image Preprocessing Techniques	Study and Implementation of Lossy Compression for Photographs Identified in the document	Experimentation on Various document images, Visualization of Result and Discussions on various finding and drawing the final Conclusions
Study and Implementation of Document Image Segmentation for separation of various regions in document	Study and Implementation of Near-lossless Compression for Text-like Regions like Figures, Tables, Equations, etc.	Thesis Writing

9. Expected Outcomes:

1. Investigation of suitable segmentation technique for document Images

2. Development of Lossless Compression technique for Textual Region

3. Implementation of Lossy Compression for Photographs in document

4. Investigation of role of Near-lossless Compression techniques considering trade-off between compression ratio and quality of reconstructed image for Text-like Regions

5. Development of the document image compression system

References:

1. Gonzalez Rafael C. and Woods Richard E., Digital Image Processing (3rd Edition), Prentice-Hall, Inc., 2006.
2. Khalid Sayood, Introduction to Data Compression (3rd Edition), Cambridge, Mogan Kaufmann, 2005, pp: 257-324.
3. A.M, Raid & Khedr, Wael & El-dosuky, Mohamed & Ahmed, Wesam. (2014). Image Compression Using Embedded ZeroTree Wavelet. Signal & Image Processing: An International Journal. 5. 33-39. 10.5121/sipij.2014.5603.
4. Kannappanavar, B., Rajanikanta, S., & Tandur, S. K. (2010). Importance of digitization of library materials. SRELS Journal of Information Management , 47 (1),113-119.
5. Salomon D, Motta G. Handbook of data compression. Springer Science & Business Media; 2010 Jan 18.
6. Martin, M.B. and Bell, A.E., 2001. New image compression techniques using multiwavelets and multiwavelet packets. IEEE Transactions on image processing, 10(4), pp.500-510.
7. Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. 2018. A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. Journal of King Saud University-Computer and Information Sciences.
8. Kasturi, R., O'gorman, L., & Govindaraju, V. 2002. Document image analysis: A Primer. Sadhana, 27(1), 3-22.
9. Doermann, D. (2014). Handbook of document image processing and recognition. K. Tombre (Ed.). New York Incorporated: Springer.
10. Shivsubramani Krishnamoorthy, R. Loganathan, K.P. Soman, "Recursive Projection Profiling for Text Image separation", *Springer;* 2010
11. Ray Smith., "Hybrid Page Layout Analysis via Tab-Stop Detection", *10th International Conference on Document Analysis and Recognition;* 2009
12. Swapnil Khedekar, Vemulapati Ramanaprasad, Srirangaraj Setlur, "Text – Image Separation in Devanagari Documents", *Proceedings of the Seventh International Conference on Document Analysis and Recognition(ICDAR'03),* 2003
13. Breuel T.M., "Two geometric algorithms for layout analysis", *Document Analysis Systems, Princeton, NJ,* 2002
14. P. Mitchell, H. Yan., "Newspaper document analysis featuring connected line segmentation", *Proceedings of International Conference on Document Analysis and Recognition, ICDAR'01,* 2001
15. Kyong-Ho Lee, Yoon-Chul Choy, Sung-Bae Cho, "Geometric Structure Analysis of Document Images: A Knowledge-Based Approach", *IEEE transactions on Pattern Analysis and Machine Intelligence,* 2000
16. Kise K., Sato A., Iwata M., "Segmentation of page images using the area Voronoi diagram", *CVIU,* 1998
17. L. Rudin, S. Osher, E. Fatemi, "Nonlinear total variation-based noise removal algorithms" *Physica D,* 1992
18. Nagy G., Seth S., Viswanathan M., "A prototype document image analysis system, for technical journals", *Computer.7,* 1992
19. Wong K.Y., Casey R.G., Wahl F.M., "Document analysis system", *IBM Journal of Research and Development,*1982

-sd-
Prashant Laxmanrao Paikrao
Research Scholar

-sd-
Dr. D. D. Doye
Supervisor
Professor, E&Tc Department
SGGSIE&T, Nanded